Artificial Intelligence in Medicine 57 (2013) 145–154

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

# A system for the extraction and representation of *summary of product characteristics* content

Stefania Rubrichi[a,*], Silvana Quaglini[a], Alex Spengler[b], Paola Russo[c], Patrick Gallinari[b]

[a] *Laboratory for Biomedical Informatics "Mario Stefanelli", Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, via Fearrata 1, 27100 Pavia, Italy*
[b] *Laboratoire d'Informatique de Paris 6, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France*
[c] *Amyloidosis Research and Treatment Center and Department of Biochemistry, IRCCS Policlinico San Matteo FDN and University of Pavia, piazzale Golgi 2, 27100 Pavia, Italy*

ARTICLE INFO

*Article history:*
Received 9 October 2011
Received in revised form 29 July 2012
Accepted 26 August 2012

*Keywords:*
Information extraction
Conditional random fields
Ontology
Summary of product characteristics
Adverse drug events
Medication errors

ABSTRACT

*Objective:* Information about medications is critical in supporting decision-making during the prescription process and thus in improving the safety and quality of care. In this work, we propose a methodology for the automatic recognition of drug-related entities (active ingredient, interaction effects, etc.) in textual drug descriptions, and their further location in a previously developed domain ontology.
*Methods and material:* The summary of product characteristics (SPC) represents the basis of information for health professionals on how to use medicines. However, this information is locked in free-text and, as such, cannot be actively accessed and elaborated by computerized applications. Our approach exploits a combination of machine learning and rule-based methods. It consists of two stages. Initially it learns to classify this information in a structured prediction framework, relying on conditional random fields. The classifier is trained and evaluated using a corpus of about a hundred SPCs. They have been hand-annotated with different semantic labels that have been derived from the domain ontology. At a second stage the extracted entities are added in the domain ontology corresponding concepts as new instances, using a set of rules manually-constructed from the corpus.
*Results:* Our evaluations show that the extraction module exhibits high overall performance, with an average F1-measure of 88% for contraindications and 90% for interactions.
*Conclusion:* SPCs can be exploited to provide structured information for computer-based decision support systems.

## 1. Introduction

The use of medications has a central role in health care provision, yet on occasion it may injure the person taking them as result of adverse drug events (ADEs). Some ADEs are inevitable and include drug effects that are unwanted, unpleasant, noxious, or potentially harmful, but sometimes they are caused by failures during the medication process, therefore they are not inevitable. Thus, to decrease injury, efforts must be put into reducing errors. Medication errors cut across multiple stages (ordering, dispensing, administration) and occur for a variety of reasons. According to a study on ADEs analysis [1], most errors occur in the ordering and administration steps: some of them are due to lack of information about the patient, but more often physicians make many prescribing errors that appear to be due to deficiencies of up-to-date knowledge of the drug and how it should be used. These include incorrect doses, forms, frequencies, and routes of administration, as well as errors in the choice of drug. Moreover, information about the patients' condition (allergies, previous diagnoses), lab results and other medicine they are taking is sometimes incomplete, not easily accessible when it is needed, which leads to prescribing errors as well as inappropriate administration of ordered drugs. Several information technologies [2,3] can help providers absorb and apply the necessary information. In particular, computerized physician order entry (CPOE) is viewed to play a key role in helping improve safety standards in the process of medication use, especially when decision support is provided. CPOE with clinical decision support can improve patients' safety and lower medication-related costs. To be effective, the underlying medical knowledge must be captured, adequately represented, and made available to the CPOE system.

In this work we consider the problem of automatic extraction of drug information conveyed in the summary of product characteristics (SPC), in the form of a formal, structured representation of their contents, suitable for a computational utilization. Therefore the overall objectives of this work fall into two tasks: the automatic recognition of drug-related entities in SPCs and the instantiation of the extracted entities in a previously developed domain ontology (i.e., ontology population). We focus on extracting information from two specific sections of SPC concerned with drug-related interactions and contraindications. Our approach combines (i) a

[*] Corresponding author. Tel.: +39 0382 985981; fax: +39 0382 525638.
*E-mail address:* stefania.rubrichi@unipv.it (S. Rubrichi).

0933-3657/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2012.08.004

rule-based method and (ii) a machine-learning method based on a state-of-the-art classifier: linear-chain conditional random fields (CRFs) classifier.

We formulate the problem of drug-related entities detection in a supervised machine-learning framework, in which we seek to assign the correct semantic label such as *InteractionEffect* or *ActiveDrugIngredient*, to each word, or segment of sentence, of the text. To this end, we introduce a corpus of about a hundred interaction and contraindication sections in the Italian language that have been annotated with thirteen distinct semantic labels, with respect to a previously implemented ontology. We apply the CRF to our data set and evaluate their overall and individual label results. The classifier achieves an average $F_1$-measure of about 90%, which is promising for real-world applications.

We then represent the extracted drug-related entities as instances of concepts in the ontology. The ontologies provide for a standardized means of modeling, querying, and reasoning over large knowledge bases. Ontology population from natural language textual document is becoming increasingly important in the domain of natural language processing (NLP) applications. We define a set of rule in order to link the output of the extraction module to the ontology. We first integrate the extracted entities according to an XML schema, then we map it into the ontology web language (OWL) format. The resulting populated ontology can then be used within any ontology-enabled system for further querying, reasoning, or other processing.

## 2. Related work

Over the last two decades there has been an increase of interest in applying NLP to biomedical text. More specifically, information extraction (IE) techniques have become a frequently-used resource for enriching the content and the utility of electronic clinical systems [4]. Excellent efforts have been documented in the literature on IE from textual clinical documents [5–9], and its subsequent application in summarization, case finding, decision-support, or statistical analysis tasks.

As far as the medications domain is concerned, several studies have addressed the issue of IE. The less recent ones concentrated their analysis on the extraction of specific drug features, such as drug names and dosage. In one example, Evans et al. [10] reported a method of extracting drug and dosage data from a collection of discharge summaries. They first draw a conceptual model of drug-dosage information and then identified this information using a semantically driven extraction module. This module combines readily available NLP facilities from the Clarit system with newly created resources, including a set of pattern rules and a lexicon. In another study, Sirohi et al. [11] performed a dictionary-based NLP study to determine the effects of using varying lexicon to extract drug names from electronic medical records. These authors have shown how the accuracy of results can be enhanced by refining the drug lexicon. A study by Shah et al. [12] derived numerical information about daily dosage from unstructured dosage instructions from a patient records database, using a dictionary to standardize words and phrases. Then they converted the extracted information into structured fields. Lastly, Levin et al. [13] implemented a system based on lexicon (RxNorm) and regular expressions (Hints List) to extract and normalize drug names from an anesthesia electronic health record, into a standardized terminology. RxNorm and Hints List concepts were used in the mapping module as references for drug names, and medical abbreviations and jargon, respectively.

Lately, more studies have been geared towards the extraction of a more complete set of drug characteristics. In particular, Gold et al. [14] built Merki, a parser which can extract drug names and other relevant information from discharge summaries using a lexicon and a set of parsing rules. Similarly, Xu et al. [15] implemented a NLP system, MedEx which extracts medication information from clinical notes. Relying on a more detailed medication representation model, they integrated a semantic tagger and Chart parser to capture drug names and signature information from clinical narratives and then to map it onto structured representation.

On the whole, current works focus on clinical text narrative. However, Pereira et al. [16] considered another source of information on medications, namely SPC, and thus addressed the problem of automatic indexing. The authors developed a method to automatically generate a dictionary for use with a French multi-terminology indexing tool. Focusing on medical literature, instead, Segura-Bedmar et al. [17] aimed at detecting drug–drug interactions by comparing two different approaches. Initially, they employed a hybrid approach, which combined shallow parsing and pattern matching, then they employed a kernel-based approach that uses support vector machines (SVM) and which achieved better results.

Broadly speaking, in the works described above, IE techniques are used for identifying specific piece of information in the text. However, IE aims also at representing the detected information in structured form that comprise meaningful association of the relevant entities. In this context, ontologies are an adequate means for modeling knowledge needed to extract and encoding data from text. Automatic ontology population from text has recently emerged as a field of application of IE where, in place of a template to be filled, the goal of the task is the extraction and classification of instances of concepts and relations defined in an ontology. In many applications, the ontology is only used to help the semantic annotation of the text using the existing concept instances. IE is then delegated to the construction of extraction patterns or rules, and thus to locate new ontological instances [18–21]. Some authors have stressed the benefit to also use the ontology as the IE output language, in the context of ontology population [22,23]. In this case, the main goal is to enrich a knowledge base with new instances specified in the ontology. Yet, the integration of ontologies in IE systems is still a research issue.

## 3. Framework outline

In this paper we propose to extract drug-related interaction and contraindication information reported as free-text in SPC and to convey them into a domain ontology. We follow a named entity recognition (NER) approach for detecting drug-related entities. NER is a sub-field of IE and refers to the task of identifying expressions in natural language text denoting certain entities (i.e., named entities), such as diseases and drugs, and labeling them with their appropriate type. To do so, we have developed a framework for simultaneously recognizing occurrences of multiple entity classes by using linear-chain CRFs. This supervised machine learning approach predicts the labels of words by using large number of interdependent descriptive characteristics (features) of the input by assigning real-value weight to these features. This can be seen as a way to “capture” the hidden patterns of labels and features, and “learn” what the likely output might be, given these patterns. Our methodology is developed through different steps.

Typically, the first step in most NER tasks is to identify the named entities (labels) that are relevant to the concepts, relations and events described in the text. A system for NER is hence based upon specific knowledge about the domain. Thus, as part of the understanding of the text factual information process, we had previously developed a formal model of drug information